IMPLEMENTATION PAPER

Implementation of the FAIR Data Principles for Exploratory Biomarker Data from Clinical Trials

Alexander Arefolov[1†], Laura Adam[1], Shoshana Brown[1], Yelena Budovskaya[1], Cong Chen[1], Diya Das[2], Chen Farhy[1], Rebecca Ferguson[1], Hongmei Huang[2], Kimberly Kanigel[1], Christina Lu[2], Oksana Polesskaya[1], Tracy Staton[3], Rajeev Tajhya[1], Maryann Whitley[1], Jee-Yeon Wong[2], Xiangpei Zeng[1] & Mark McCreary[2]

[1]Rancho BioSciences LLC., San Diego, CA 92127, USA

[2]Development Sciences Informatics, Genentech Inc., South San Francisco, CA 94080-4990, USA

[3]Development Sciences OMNI-Biomarker Development, Genentech Inc., South San Francisco, CA 94080-4990, USA

Keywords: Data management; Data curation; Clinical studies; FAIR data; Biomarkers; Data model; Metadata; Data standards; Data governance

Citation: Arefolov, A., et al.: Implementation of the FAIR data principles for exploratory biomarker data from clinical trials. Data Intelligence 3(4), 630-661 (2021). doi: 10.1162/dint_a_00106

Received: April 28, 2021; Revised: June 29, 2021; Accepted: July 6, 2021

ABSTRACT

The FAIR data guiding principles have been recently developed and widely adopted to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets in the face of an exponential increase of data volume and complexity. The FAIR data principles have been formulated on a general level and the technological implementation of these principles remains up to the industries and organizations working on maximizing the value of their data. Here, we describe the data management and curation methodologies and best practices developed for FAIRification of clinical exploratory biomarker data collected from over 250 clinical studies. We discuss the data curation effort involved, the resulting output, and the business and scientific impact of our work. Finally, we propose prospective planning for FAIR data to optimize data management efforts and maximize data value.

1. INTRODUCTION

As the pharmaceutical industry evolves to combine the latest advances of medical knowledge and technology, the amount and complexity of healthcare data is increasing exponentially with an estimated

[†] Corresponding author: Alexander Arefolov (Email: alex.arefolov@ranchobiosciences.com; ORCID: 0000-0002-3447-3223).

compound annual growth rate of 36% [1]. Healthcare data volume has reached the zettabyte scale by some projections [2] with approximately 80% of the data remaining unstructured and poorly organized [3], thus of limited utility for downstream analysis without extensive data curation efforts.

To address the challenges of data management and stewardship across all data-intensive industries, the FAIR (Findable, Accessible, Interoperable, and Reusable) data guiding principles have been recently developed by a diverse set of academic, corporate and governmental stakeholders [4, 5]. These principles have been quickly adopted by publishers, funding agencies and international intergovernmental organizations [6, 7, 8]. Successful implementation of the FAIR principles by the pharmaceutical industry is a key prerequisite for digital transformation [9, 10] and the disruptive potential of machine learning and artificial intelligence in drug discovery [11]. Additionally, it is highly anticipated that data FAIRification will accelerate innovation, aid in the development of personalized medicine, drive down drug development timelines, reduce R&D costs and enable data sharing between research groups within and across institutions and companies [12].

While still in the early stages, industry-wide implementation of the FAIR data guiding principles is faced with many challenges, including the costs of FAIRification, limited understanding and availability of technology and standards to support FAIR implementation, and the need for cultural change within the organizations. Addressing some of those challenges, a bottom-up approach for data integration driven by use cases would allow the organization to begin building the basic infrastructure and standards to make data FAIR, increase organizational awareness, demonstrate business impact and build capabilities for the future.

Among the different types of healthcare data, biomarker data have played an increasingly significant role in drug development in recent decades [13, 14, 15, 16]. Biomarkers are used to understand mechanisms of action, evaluate pharmacology (pharmacokinetics/ pharmacodynamics), explain differences in treatment responses, and select or stratify subjects. A study of clinical development success rates from 2006 to 2015 demonstrated benefits of using biomarkers for subject selection alone—a threefold increase in the probability of approval from Phase I and 20% increase in the transition to approval from Phase III [16]. While essential for utilizing its full value in drug development, FAIRification of biomarker data has its own set of challenges. First, there is a broad and continuously evolving diversity of assays—from single immunoassays to complex multicolor flow cytometry. Second, assay workflows are diverse as each assay type has unique workflow and quality control parameters. Third, clinical sample flow is complex as different specimen types are taken at multiple clinical sites, and then the specimens (and subsequently derived samples) are processed and sent from the first vendor to downstream vendor #2 and so on, often resulting in a convoluted sample flow chart. Fourth, some therapeutic areas, like oncology, are especially complex. Disease pathogenesis and heterogeneity, drug response and resistance, genetics and the impact of the immune system, among other factors, each needs to be researched and carefully considered to effectively address the complexity of the disease, creating a different type of data complexity in the process.

Traditional workflows have been inadequate for handling the data volume and the complexity of biomarker data harmonization and FAIRification, often resulting in significant delays, or even reduced data availability for downstream analysis and lost value altogether. Creating new and more efficient workflows for biomarker data FAIRification and integration is therefore crucial to enable the data to be used for future exploration of scientific questions and data analysis, meeting regulatory requirements and supporting go/no go decisions at each stage of clinical trial.

This paper focuses on the methodology and best practices for **data curation and metadata mapping** developed for FAIRification of clinical exploratory biomarker data from several use cases. The biomarker data were collected from over 250 legacy studies (defined by the FDA as a set of data that is not developed based on current FDA-endorsed data standards), such as the CDISC SDTM data model [17, 18] and on-going clinical trials and contained a wide variety of data types including next-generation sequencing (NGS), immunohistochemistry, flow cytometry, imaging and mass spectrometry, among other assay types.

Once the necessary models and processes are in place, a data repository can be developed to share and explore the data to enable scientists to submit queries, such as in the following scenarios:

- For hypothesis exploration, identify all data sets that include "my favorite gene" and report assay and study metadata to allow selection of data set(s);
- For expression quantitative trait loci analysis (eQTL) in a particular disease, identify paired whole exome sequencing (WES) and RNA-Seq samples originating from the same specimens;
- For meta-analysis of treatment effect, identify pre- and post-treatment RNA-Seq fastq files that meet provided quality control (QC) criteria and group by laboratory batch.

While the details of the final implementation may vary depending on user needs, the process described in the following sections should remain the same. This paper will share insights and best data management practices from our experience with a large and varied biomarker collection and conclude with a discussion of future plans.

2. CONCEPTUAL DATA MODEL FOR BIOMARKER METADATA MAPPING

Harmonized and linked metadata is a powerful tool for data stewardship as it enables and supports all four key principles of FAIR data [4]. One of the crucial steps in the data FAIRification workflow is to establish a metadata model that allows efficient data integration and complex queries in the final data repository. This is a key element of an open-data ecosystem in which data are valued, preserved and reused. As a full data model is out of scope for this work and has been proposed by others [19, 20, 21], we will focus on the aspects minimally required as a framework for data FAIRification. The model presented here is focused on sufficient metadata to uniquely identify subjects, samples and assays and was developed de novo after several iterations. With these in place, links can then be created and maintained between subject, sample, assay and data, thus creating a framework that can be supplemented with additional metadata or linked to other source systems.

2.1 Study Subject

In the clinical biomarker data environment, a clinical repository is usually the source system for subject, treatment, and disease characteristics. As the goal of a minimal metadata model is to maintain sufficient metadata to link to source systems where available, in the case of clinical trial data, it is not necessary to maintain subject-level metadata (e.g., age, sex, disease) with the biomarker results when a small number of identifiers are sufficient to link to the rich source of information in the clinical trial record. Therefore, the focus of the FAIRication effort for subjects is to map the study identifiers and subject identifiers assigned to the biomarker data to the identifiers in the clinical trial records. This step ensures the validity of data through downstream anonymization processes that are often required for re-use or sharing of the data.

While this may appear to be simple, in reality, these identifiers can take different forms over the life of the trial including the use of study names or aliases in place of the unique study identifier, and the re-use of subject identifiers from one study to the next.

2.2 Specimen and Derived Samples

Tracing sample identifiers reported with biomarker data to the original specimen and subject is one of the most time-consuming aspects of data integration. The goal of FAIRification is to capture the sample metadata required to uniquely define a sample and to link it to the subject from which it was collected. In our case, there was no sample identifier source system that could provide the lineage for all sample identifiers, resulting in the need to build that lineage as part of the data curation and mapping process. The degree to which the clinical data repository contains biomarker sample metadata varies greatly across the industry as well as from study to study within one organization. While the identifier initially assigned to the specimen when it is taken from the subject is sometimes recorded in the clinical record, this is not a universal practice. Furthermore, as the sample is processed, new identifiers are generated. Sample derivation lineage is an important aspect of sample metadata, supporting the need to distinguish technical data duplication from data produced from unique samples collected simultaneously. The lineage sample identifiers trace a sample minimally to the immediate parent sample. If sample processing practices allow, it is optimal to also retain the original specimen identifier, the date of sample collection, the trial visit identifier, and the anatomical location from which the specimen was collected. These allow for a robust integration with the clinical record. Related to the downstream-derived samples, we record the type of the sample and organization that produced it. Lineage sample identifiers and the minimal metadata are shown in Figure 1.

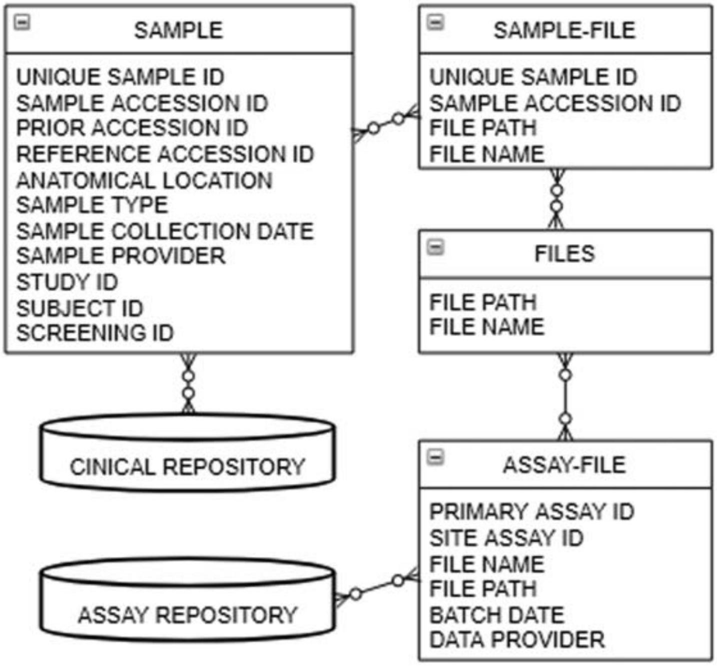

Figure 1. FAIRification of exploratory biomarker data by linking metadata for subjects, samples, assays and data. This minimal **logical** model of metadata and relationships is a visual representation of the curation outputs.

2.3 Assays

Biomarker assays are routinely updated or replaced by newer technology. The minimal metadata for assays must capture the differences between related but non-comparable assays. Due to the complex variables that define an assay, it is optimal to store assay data in a repository that can serve as a source system for linking the data files to the assay definitions. When this work began, an assay repository was available but most biomarker assays were not recorded. The curation effort included obtaining assay details to enter into the assay repository. While the full model of the assay repository is out of scope for this work, we will describe the process of capturing the required metadata to assign the unique assay identifier required for the FAIRification process.

2.4 Relationships

Once subjects and samples are mapped and assays are defined, the final step is to create the relationships. The data files are the products of an assay performed on a sample. The sample identifier and the assay identifiers are thus linked to the data file identifiers, as shown conceptually in Figure 1B. In operation, the curated tables are generated with file names and file paths, which are used to determine the appropriate unique identifier for the file in the biomarker repository. Curation processes for each of these tables are discussed in greater detail below.

3. LINKING STUDY SUBJECTS AND SAMPLES

The FAIRification process entails establishing connections between the study subjects and samples taken from them as well as relevant metadata for each sample (e.g., sample type, anatomical location, processing vendor, collection date-time). Clinical trial biological samples may be obtained when screening subjects pre-enrollment as well as during and post treatment. In general, clinical sites send samples to a central laboratory for storage and processing. Central labs then ship samples to 'downstream' vendors. These vendors execute the assays and may further process the samples. Vendors may also be asked to send samples to another laboratory downstream. Therefore, a key aspect of the process is tracing the lineage of sample derivation beginning with its collection at the study site. However, the lack of standards for maintaining linkages between sample, parent sample, subject identifier and sample metadata renders this task labor-intensive and resistant to automation. In this section, we describe our process for tracing and maintaining the linkage of study subjects and samples.

In this context, we find it convenient to discuss differences between high-dimensional and low-dimensional data harmonization cases. High-dimensional data are categorized here as data generated by standardized assays involving high-throughput methodologies, most commonly omics data such as various uses of next-generation sequencing. High-dimensional biomarker data sets typically consist of an extremely large number of potential markers measured in relatively few samples of subjects. As these methods are generally newer and designed for automated downstream analysis, the data organization and formatting is usually coherent and consistent. Low-dimensional data are categorized here as the cases where results are collected into a data summary file. These aggregate results often incorporate results for multiple assays together and are common for low-throughput methods, such as immunoassays (e.g., ELISA, PCR, and some histology assays). These summary data files present their own set of challenges for FAIRification.

3.1 Input Data

Collaboration with the biomarker operations component of the study teams is critical to access the required input data for the linking work. A vital input in the Subject-Sample mapping process is the biosample management plan. This is usually presented as a graphic flowchart in a study planning document and conveys the samples' processing history. Subject-Sample data are usually shown as tabular manifest files containing a combination of subject identification, sample identification and reference accession identification and at times additional sample metadata (e.g., type or date). For high-dimensional data, these files are usually processed from each run where each row contains data for a single sample. However, sample metadata including subject, visit, or sample type is not readily available from vendor-provided data, which hinders linkage efforts. Inputs for low-dimensional data are often assay results files. These are generally easier to map to subject and sample identifiers as much of the necessary information is present within the file itself. However, there can be more variation in file format than in high-dimensional data, making it more difficult to programmatically merge and analyze data even within a single study. Thus, the main challenges lie in integrating assay metadata and processing files into a standard format that can be easily merged and analyzed.

3.2 Challenges

3.2.1 Legacy Data

The lack of business continuity is a major challenge in curating legacy studies. No central repository nor single owner exists, which hinders access to required information. Closed studies may have out-of-date legacy documentation, or documents may be lost altogether when laboratories close, people move and antiquated data management systems are phased out. Hence, metadata for legacy studies may not be readily available, requiring extensive inventory of available resources.

3.2.2 Ownership Changes Throughout Clinical Data Lifecycle

For both high- and low-dimensional data, documentation of the complete lifecycle of the samples is crucial to track a sample identifier from an assay back to subject-level information. With high-dimensional data, laboratories that perform assays and generate the data often have limited sample information, and necessary sample metadata (e.g., subject or visit information) which is commonly missing in the tables used to reconstruct a sample's lifecycle.

3.2.3 Lack of Standardization

For high-dimensional data, sample metadata is often found in manifest files, which are non-standard and therefore difficult to integrate. Manifests are generally manually edited, for instance when samples are combined to extract enough DNA. As a result, they contain errors, especially since files are frequently provided as Excel spreadsheets which can give rise to formatting errors. In addition, it is not uncommon for a variable to contain a mix of values from two different sources (e.g., the screening and enrollment identifiers). We have even encountered cases where assay and reference accession identifications are hard-coded in raw data file names instead of being present in a manifest.

Low-dimensional results files come in a large variety of types, including delimited text, Excel, and SAS formats. Automated curation scripts must be able to determine the file type to appropriately parse the input file. Regardless of file type, data organization can vary greatly from assay to assay or from vendor to vendor. A parsed data format is a common example of such inconsistency. While some files follow a tall-skinny format, wherein data from a single subject, test, or visit are found in their own row, other files may adhere to a wide format, where results from different tests or visits are split across variables. Thus, automated scripts must be able to standardize both tall-skinny and wide data into a single standard format for further processing. Inconsistencies with variable name and formatting can also occur in both low- and high-dimensional data. In some files, variable names are specified on the first row of the input file, and the file contains no irrelevant information that must be removed. In other cases, variable names span multiple rows or may not begin on the first row of the file. In addition, names for variables that represent the same information may differ across files, so automated scripts must be able to standardize variable names for important variables to be recognized and processed appropriately. Furthermore, the data themselves may be reported in different formats: date values may not be formatted to a common standard and text values representing the same concept may be specified differently. Standardization of each type of data may require a separate process.

3.2.4 Atypical Vendors

Atypical vendors, such as academic laboratories, may provide primary- and meta-data requiring custom scripts or manual curation. For example, we have encountered tissue microarray slide data files described so informally in a spreadsheet that interpretation by the curator was required to determine the tissue's source.

3.2.5 Non-tabular Data

Finally, some information about the samples is not contained in table form, but is found in reference documents, such as sample origin or visit codes, especially when those data are the same across all study samples.

3.3 Process

To link study subjects and samples, we build the chain of sample identifiers all the way back to the central lab that holds a manifest file containing all the metadata. For each sample, we collect all subject identifiers, sample identifiers, reference accession identifications and lineage information, sample type, anatomic location, date and time of sample collection, visit information and vendor. The Subject-Sample mapping process starts with surveying the study data and identifying available sources of metadata. First, we locate and manually parse the biosample management plan. We identify the type of identifier used and the documentation available for each assay and vendor. If the samples do not originate in a central lab, we locate the shipment file that maps the assay sample identifiers to the previous lab identifiers. Despite all efforts, some information may still be missing. Ongoing studies have a point of contact from whom we can request metadata files that are not readily available, or who can clarify the sample flow. In the case of legacy studies, missing data may be irredeemably lost and the curator often must decide how to handle each case.

Many aspects of the curation process can be automated. However, the wide variety of input file types and formats renders it difficult for complete automation. Thus, we use a hybrid approach in which we intersperse prewritten functions or code chunks that perform common curation tasks with custom code required to process a specific data set into the format required by these functions. While parsing available documents (e.g., manifests and assay results files) we find that many contain similar types of information. However, variable naming conventions may differ significantly. Identifying relevant variables, renaming them to convention and removing unnecessary variables can be mostly automated, given a dictionary that specifies common variable names and their standardized counterparts.

Once all available sample information is collected and checked by the curator, the table undergoes QC.

3.4 Quality Control

To check the quality of the output table produced, we established a QC pipeline that includes both scripted and human aspects. The automated QC script consists of over 50 checks, including: (1) Ensuring fields conform to the appropriate controlled vocabulary, such as anatomical location and vendor names; (2) Ensuring date fields are in the correct format; (3) Ensuring all required data are provided; (4) Ensuring fields are properly ordered; (5) Ensuring samples have a unique origin; and (6) Ensuring attributes (e.g., date and visit) are consistent with those of the downstream samples. For manual curation, a second curator (usually more experienced) interprets the results of the QC script, traces the sources of any errors to allow for exceptions, validates the sample flow chart interpretation and performs a visual inspection of the curated output file and the curation script to identify any potential issues missed by automated QC. Any problems identified during the QC process are corrected by modifying the curation script, and a final version of the curated output file is generated.

3.5 Output

The curated output is written to a standard .csv file. This file contains standard variable names specified in a standard order (Figure 1). Data within a given field are standardized according to that field's respective rules. Each row contains a sample identifier, the direct reference accession identification, the central lab reference accession identification (if it exists) and sample metadata.

4. COLLECTING AND STORING ASSAY METADATA

FAIRification requires a searchable central assay repository that stores sufficient metadata to distinguish between related but non-identical assays and assigns a unique identifier to each assay. Many considerations behind establishing such a database, like the underlying software component (storage engine) behind its management system, are beyond the scope of this paper. One critical component is the level of detail the database provides, which would eventually set the limitations on distinguishing assays. These considerations must take into account not only the type and diversity of assays that would be registered in the database but also the information expected to be available to the curators. The assay repository employed here holds a large number of distinct fields. These include (Supplementary Table 1) information regarding the purpose of the assay, the methodology, platform and vendor (e.g., ELISA, Simoa HD-1 Analyzer, Myriad RBM), analyte examined (e.g., type, gene symbol, UniProt ID), the acceptable tissue or specimen (e.g., plasma, tumor, normal) and technical details (e.g., quantitative limits and units). With this level of detail, curators can coherently and consistently define assays across studies. Harmonization of the terminologies (such as gene and vendor names) allows curators to determine whether a performed assay should be recorded as a new assay, or whether it is already recorded in the repository. More importantly, downstream users can generate informative queries across multiple studies and projects. Once established, the platform is manually populated by teams of curators that scan the available documentation associated with each data set collected, and register the assays used. The following subsections discuss available sources of assay information, the process of deriving it, and challenges and workarounds developed by curators.

4.1 Input/Challenges

The process for assay registration entails collecting assay information from various sources (detailed below) and recording it in a standardized manner. A key issue in this process is determining what constitutes an assay, and which parameters and metadata are sufficient for distinguishing one assay from the other. The curation process is designed to create and record this metadata and address challenges such as incomplete information, multiple sources and formats, and lack of standard vocabularies as detailed below.

4.2 Input Data

Input data for collecting and organizing metadata for assays can be found in several types of documents. While some of these, such as validated assay protocols, contain all the required information, but others, such as sample transfer documentation, typically contain only partial information. Table 1 below details some of the information required to uniquely distinguish assays and where to find each piece of information. If these sources are unavailable or incomplete, curators can attempt to contact the scientist commissioning these assays internally, or the vendor directly. In many cases, the curation team was able to develop excellent collaborations with internal biomarker scientists to support the assay work. Despite this wealth of sources, curators still encounter several common challenges.

Table 1. Information required to uniquely distinguish assays.

		Validated assay protocols	Study protocols	Sample transfer documentation	Commercial vendors' websites	Internal databases	Published papers on the studies	Vendor or scientist
Assay information	Name	Yes			Yes			
	Vendor	Yes	Yes	Yes				Yes
Methodology	General Method	Yes	Some		Some	Some	Yes	Yes
	Sub-Method	Yes	Some		Some	Some	Yes	Yes
	Detection Method							
	Detection/measurement device	Yes					Some	Some
Biomarker/ Analyte	Universal Identifier (e.g., UniProtID)	Some			Some	Some		
	Analyte/Panel name	Yes	Some	Yes	Yes	Yes	Yes	Yes
	Analyte type (e.g., Protein)	Yes	Some	Some	Yes	Some	Yes	Yes
	Measurement category	Yes	Some		Some	Some	Some	Yes
Technical details	Detection reagent (e.g., Antibody clone)	Yes	Some		Some	Some	Some	Yes
	Limits of quantification	Yes			Some	Some	Some	Yes
	Measurement unit	Yes	Some		Yes	Some	Yes	Yes
Tissue of interest	Species	Yes	Yes		Yes	Yes	Yes	Yes
	Biological matrix	Yes	Yes	Some	Some	Some	Some	Yes

4.3 Challenges

4.3.1 Legacy Studies or Partial Information

Missing data is common in studies conducted prior to the data explosion era as well as in newer studies which do not employ modern data management practices. In these studies, assay information is often poorly annotated. If enough time has elapsed and the methods are outdated, records may be lost and the personnel performing the assays (and sometimes the vendors) can no longer be reached. In such cases, a curator will attempt to derive the information from all available sources. However, despite all efforts, legacy studies commonly have only partial assay information, rendering it impossible to determine the exact assay used. Such endeavors cannot be automated and require manual curation by a scientist who is an expert in the field.

4.3.2 Standardized Vocabulary

Data analysis automation efforts rely on compact, non-duplicative, terminology standardized across all essential elements of assay information. These standardized terminologies, or code lists, minimize the variations and nuances found in natural language, thus supporting scripted data wrangling efforts. For example, the "Hematoxylin-Eosin" assay is referred to as "HE," "H&E," or other variation of its full name. Adhering to standardized names enables easy collection of data generated with this assay across all curated studies. While codification of common assays, as well as many analytes such as proteins or genes, is straightforward and relies on universal identifiers such as UniProt IDs or HUGO IDs, other assays or biomarkers are more challenging. For example, the terminology describing sorted cell subpopulations that express a particular set of surface marker combinations is constantly evolving and has not been affected by global efforts of harmonization. Consequently, code lists for assay details contain both universal and local codes. The assay metadata curation effort therefore requires constant maintenance of code lists that ensure the use of universal identifiers whenever possible and are shared between database maintainers and users.

4.3.3 Distinguishing Closely Related Assays

This issue stems from the broad and ever-growing variety of biomarker assays which is a direct result of both technological progress and proliferation of vendors endeavoring to fine-tune each assay to exact specifications. This heterogeneity spans multiple aspects, including differences in antibody clones for histological analysis, kits used to prepare RNA libraries for sequencing and platforms on which the libraries are run. This vast diversity means that instead of discrete, easily distinguishable assays, curators are often faced with a spectrum of closely related assays. We have developed a set of rules to group or distinguish between seemingly similar assays. For example: (1) Next-generation sequencing uses a specific library preparation method, and then this library is sequenced. Sequencing can run for a number of cycles, resulting in different read lengths. We consider assays that used different read lengths as the same method, as long as library preparation is the same; and (2) Immunoassays that use different monoclonal antibodies to the same antigen are different assays, even if all other parameters of the assay are identical.

4.4 Process

There are multiple stages in the process of collecting metadata when registering an assay (creating a new record in the database). Curators first explore all available information sources, allowing them to next identify the analyte (in standardized terminology, such as UniProt ID). Older studies that use outdated protein or gene names or high-throughput assays that test multiple analytes may present complications. If assays have a large number of analytes, the analyte list is compared programmatically with similar, previously recorded assays to attempt a match. If no exact match is found, the curator must determine whether this is a new assay or a subset of the analytes that has been reported from an existing assay. For assays that do not have a list of analytes, such as whole genome sequencing which uses library preparation methods based on random priming, assays are determined by method and protocol only.

Next, the curator identifies the vendor, after which he/she can inquire if an assay for the analyte already exists within the repository. If yes, metadata for the registered assay is compared to all available data protocols to confirm the assay is indeed the same. If the registered assay has a different vendor, a new "site assay" can be created with vendor-specific parameters (e.g., quantification limits). If a vendor cannot be identified, the curator must gather information about the assay and prepare it for registration. All required fields (e.g., general method, submethod, biological matrix, measurement units; see Supplementary Table 1) should be identified. If information is not readily available from existing documentation, curators then research literature or reach out to the vendor or scientist that performed the assay. Documentation must include how information was obtained, complete with reference papers or correspondences. Then, a concise and informative assay description may be created.

After all information is gathered into the repository, a new record is created. The curator executes a QC check, and then sends the record to a second curator for an additional QC check. After all issues are resolved, the final iteration is submitted to the repository as a new assay. Finally, once the supervisor performs a cursory check and approves the assay, an assay identifier is assigned.

4.5 Quality Control

Due to format differences and source availability between studies, the assays are manually registered by specialized curators with the required domain knowledge. All decisions made are then QC-ed by a second specialized curator. The QC curator reviews all available sources submitted by the first curator, including communications with internal scientists or the vendor, and compares these to the information entered into the database. Programmatic QC steps are used to confirm analytes lists in assays with a large number of analytes.

4.6 Output

The outcome of the assay FAIRification process is a central repository containing all assays harmonized across studies, with a rich metadata layer. Downstream, these efforts ensure that programmatic data sectioning across studies considers differences in methodological approaches to determine appropriate covariates, decreasing data noise and artifacts. However, before this occurs, the registered assays need to be associated with individual data files which are in turn associated with individual samples and subjects.

5. LINKING SAMPLES, DATA FILES AND ASSAYS

Data FAIRification entails transferring data from multiple studies and sources into a central location. Large organizations or groups may have legacy data stored across multiple cloud sources and individual computers, as well as incoming data from vendors. If data are transferred to a single storage site (such as a centralized file repository) it can be accessed organization-wide. It is then given information-rich metadata labels, allowing fast searches for data subsets. All files are cataloged and associated with the samples for which they were derived and with metadata regarding the assay used to generate them. Ideally, this association is concurrent with the data transfer from the vendor. However, metadata is currently assigned after the data have already been transferred and stored. In this section, we describe the process for creating both the sample-to-file and file-to-assay associations.

5.1 Assigning Assay Metadata to Data Files

Incoming data are initially organized by a receiving specialist or a curator group specializing in biomarkers. These curators receive the data files and conduct a superficial mining of the data being transferred to detect general information, such as the overarching technology used (e.g., histology, FACS, next generation sequencing) and place the data files in a hierarchical directory consistent across studies and assays. These are supplemented by vendor tags, resulting in a name and path for each data file that contains some assay information. The following chapter details the procedure we developed to map assay labels (assay identifications) and other assay metadata to data files based on existing data structure and available information. This process includes manual and scripted components, and multiple rounds of organization and identification to label a large number of data files accurately and efficiently.

5.1.2 Inputs/Challenges

The primary inputs for file-to-assay association are the data file names and paths tagged by vendors and receiving scientists to include rudimentary assay information. These are treated as strings to be parsed to pull metadata. As secondary input into this process, we employ elaborate dictionaries consisting of several hundred keywords that link particular substrings within the file path with known files subtypes (e.g., readme, manifest), vendors and assays. This approach underwent several rounds of development. Initially, specialized curators manually associated files with a particular assay following assay verification or registration (see Section 4). Once we had a sufficient number of assays and associated files, we manually generated keywords to reliably associate information, such as vendor and the assay, with each file. These keyword dictionaries were repeatedly updated, becoming more elaborate and precise. The manual approach for generating these keywords reflects the large differences between naming conventions across studies or vendors, as well as the lower number of files and assays curated during early stages which precluded a more systemic, machine learning-based approach.

To estimate the accuracy of the keyword dictionary approach across different technologies, we conducted a retrospective analysis comparing the automated assay assignment with post-curation assignment. The results (not shown) are highly technology-dependent. Some assays, such as the next-generation sequencing assays often achieve over 95% accuracy as these are usually run on highly automated modern platforms that generate a well-structured and consistent set of output files meant for scripted analysis. However, this approach faces some challenges when dealing with legacy data or methods for which information-rich standard output structure has not been established. The primary challenges we encounter are file names and paths that lack descriptive information or the available keywords in the file path are not assay-specific (as is the case with flow cytometry where the average accuracy is only 9%)

Another issue is fine-tuning keyword precision (defined as the ratio of true positives to total predicted positives). From the standpoint of curation time and efforts, correcting false positive assignment is as time consuming as manual assignment. We therefore skewed our dictionary towards longer keywords for higher true positives over false positive ratio. For instance, as mentioned before, the common histologic assay Hematoxylin-Eosin stain can be abbreviated as "HE," "H&E," or take a similar form. Two-letter combinations are highly prone to false positive detection and are found in many other words, including drug and sample names. However, when we increase keyword length, we also increase false negative detection. Optimally, multiple longer keyword permutations should be included. For example, "HE" as it refers to the assay often appears as "_HE_" but can also appear in similar permutations such as "_HE," "HE_," and "-HE-," all of which we include as keywords associated with the same assay. Taking histology files as example, the result of keyword optimization can be seen in the accuracy analysis (data not shown) showing about 50% true positive and about 50% no match with negligible number of false positives. As we accumulate more curated data that can be used to train classification models, we intend to abandon the keywords in favor of a machine learning based approach.

5.1.3 Process

Curation begins with a series of simple scripted steps. As all files pertaining to a particular study are already grouped into a compact folder structure, it is programmatically simple to generate a complete list of files and curator input is not required. Each file listed is tagged with previously known information (e.g., study name, the server housing the files). In the next step, file paths from each technology (e.g., ngs, histology) are scanned against a specialized set of keywords and regular expressions to derive file extensions, file subtypes, batch date, vendor and preliminary primary assay identification. This results in a table that lists all files in the study path as well as their programmatically derived metadata.

Files related to each assay are tagged for "data level." Data produced directly are machine output and recorded observation is "level 1." Processed data, such as bioinformatics analysis pipeline output, is designated as "level 2." Finally, organizational documents, such as sample manifests and sample flow plans, contain no data or derivatives but can be necessary for interpreting or labeling the data. These are labeled as "level 3." Though we derive metadata for all, only files considered primary data, or "level 1," are assigned assay metadata. This process is done programmatically based on file extension and subtype to add a level tag to the various file categories, such as those created automatically by the operating system (such as thumbs.db) and documentation files (such as readme.pdf).

This table is then reviewed and, if necessary, edited by a curator to ensure information such as batch date, vendor and primary assay identification were correctly assigned. Primary assay identification and vendor combinations are unique, and once both are correctly assigned, a script automatically adds the right site assay identification to the table. It is during this manual validation process that curators review documents containing assay metadata (see Section 4.2) and register any assays missing from the repository. From experience, we have learned that for some technologies, such as next-generation sequencing techniques, all files within a single folder are associated with the same assay. Thus, this validation of the derived metadata is done on a per-folder basis. This represents a significant reduction in curation efforts as the number of folders can be several orders of magnitudes lower than the number of files. Unfortunately, for technologies such as histology, folders often contain files associated with different assays and require a time consuming per-file validation process. The finalized validated table is then queued for QC.

5.1.4 Quality Control

The file metadata table undergoes both manual and scripted QC steps. During manual QC, a second curator goes over the table and the associated assay documentation. Automated QC involves over 30 scripted checks. These include flagging missing values, testing for incorrect format, confirming that all files in the list exist on the server and that all files on the server are accounted for in the list. Most importantly, the script cross-references the derived data with vendor code list and with the assay registry.

5.1.5 Output

The final table contains metadata including assay identification, vendor, batch date, file extension and subtype, for all files containing primary data (Figure. 1).

5.2 Linking Samples to Data Files

To complete the linkage of metadata for subjects-samples-assays, these objects should be mapped to the data files associated with them. The rest of this section will describe the process of mapping samples to files associated with those samples, highlighting the differences between mapping high-dimensional and low-dimensional data.

5.2.1 Inputs/Challenges

After the data have been transferred and organized into the centralized fileshare location for each study, the file path serves as the main input that allows the curators to programmatically find all files containing sample identifiers within the study folder. In addition, the output table mapping subjects to samples (see Section 3.5) serves as a secondary input that allows identification of samples that were not linked to subjects for QC purposes.

While the process of extracting the sample identifiers from files is conceptually simple and can be mostly automated, major challenges in sample-to-file mapping include variability in types and formats of files containing sample identifiers, inconsistency in variable names containing the sample identifiers within these files, and differences and non-uniqueness in alphanumeric formats for the sample identifiers themselves.

5.2.2 Different File Types and File Formats

For low-dimensional data where the results for multiple samples are typically collected into data summary files, the most common file types are text and delimited text, Excel and SAS. These files typically contain tabular data in a wide variety of formats. Besides the preferable "tidy" data format [22] where each row belongs to a unique sample and sample identifiers can be found in just one of the variables, we have encountered multiple other file formats. Some examples include files with more than one variable per sample identifier, where the data table is preceded by a header of several rows, sample identifiers can be contained in columns or in rows, and Excel files that contain multiple worksheets (with sample identifiers contained in one or several of those worksheets). These variations make it challenging to fully automate sample identifier extraction from all data files, as a curator must intervene when a file with a new format appears that could not be processed by a script.

High-dimensional data are often communicated by vendors as one file per sample per run. For files containing high-dimensional data, sample identification is usually embedded as part of the file name. For this type of file, successful extraction of the sample identifier depends on how easily a sample identifier can be distinguished from the other alphanumeric characters in the file name.

5.2.3 Variability in Column Names Containing Sample Identifiers

For the low-dimensional files, automation of sample identifier extraction depends on reliable identification of the columns (or rows) that contain the sample identifiers. In practice, however, sample identifier column names can vary. While the script can choose a column name from an array of all possible names encountered in previous studies, sometimes a new column name could result in a subset of files that cannot be processed by standardized script and must be processed separately.

5.2.4 Non-unique Formats for Sample Identifiers

Lack of format standardization for sample identifiers can result in non-unique simplistic formats that lead to difficulties with sample identifier extraction from the file names. For high-dimensional data where sample identifiers are part of the file name, scripts rely on regular expressions to distinguish and extract the sample identifier from the rest of the file name. If the sample identifier format is not sufficiently unique, regular expressions may not be able to distinguish between the sample identifier and the rest of the file name.

5.2.5 Duplicated Sample Identifiers

Ideally, every sample identifier should describe a unique sample. However, there are instances when Subject-Sample mapping output contains duplicated sample identifiers. Since the sample identifiers associate files with samples during the sample-to-file mapping process, samples with non-unique identifiers are linked with their data files as well as the data files related to samples with shared identifiers. This generates a cascade of errors not only in sample-to-file mapping but also in sample-to-assay mapping. This sometimes results from poor sample management practices when the same sample identifier is used for parent and daughter samples by the data providers.

5.3 Process

The process of sample-to-file mapping consists of running standardized scripts to extract sample identifiers from file names in high-dimensional data or read in the files, extract the sample identifiers from low-dimensional data to produce tabular output with file names, file paths and sample identifiers. Separate standardized scripts were created for every technology (e.g., ngs, histology) to accommodate file type and format variability. Part of the design process for the standardized scripts was to incorporate feedback from every new study mapped to include new variable names and formats for the files for sample identifier extraction with fewer errors.

Once the standardized scripts are run for every technology, the curator must check that sample identifiers were extracted correctly and that all the files containing sample identifiers have been processed by the scripts. If some data files could not be processed by the scripts (for example, due to an unexpected file format or a new name for a variable containing sample identifiers), the curators add the necessary code to complete the sample-to-file mapping process and correct any errors.

When non-unique sample identifiers are present in the Subject-Sample output, it could be necessary to include additional variables to the sample-file table to uniquely identify the samples.

5.4 Quality Control

Quality control checks, both manual and scripted, are incorporated into the sample-to-file mapping process. After the standardized scripts produce the tabular output containing file paths, file names and sample identifiers extracted from these files (or sample identifiers extracted from the file names in case of high-dimensional data), the curator visually checks the extracted sample identifiers for errors. Following that step, the curator runs a QC script comparing the list of data files from the output to the total list of data files in the file repository and reporting the difference. The curator then goes through the list of unmapped files to see if these files contain sample identifiers. If unmapped files contain sample identifiers, the curator adds the necessary code to read in those files, extract the sample identifiers and add the data to the output. Next, the curator runs another QC script that compares the extracted sample identifiers to the sample identifiers from the Subject-Sample mapping process described in Section 3. This comparison identifies sample identifiers missed by the Subject-Sample mapping process and catches incorrectly extracted sample identifiers. Following this verification, only the rows with sample identifiers found in the Subject-Sample mapping process remain in the sample-to-file mapping output. Finally, another curator performs a QC check ensuring all steps were followed correctly.

5.5 Output

The curated output is written to a file in standard .csv format. This file contains all the file paths and names for the primary data and sample identifiers extracted from those files (Figure 1). Coupled with two additional tables, one which records sample metadata and one linking assay metadata to primary data files, these tables allow unique mapping of all samples to standardized assays and vice versa.

6. PLANNING FOR FAIR DATA

With our previous experience in mind, the goal of the future process is to include FAIR data requirements at the time of data generation planning. This has several advantages to the retrospective method. (1) time to prospectively plan is far less than time needed to retrospectively curate; (2) metadata requirements can be determined in advance for a given data type and applied consistently to different sources of the same data type; and (3) data format standardization allows for parsing and/or processing of the data through pipelines programmatically, further reducing the time and effort needed to deliver the data to scientists and analysts.

In general, the planning phase should include the data management team, the biomarker scientist(s), and a metadata expert. These three components work together to assess the biomarker needs, select specific assays, define the laboratories that will generate the data, define the required metadata and communicate requirements to the laboratories that will generate the data. A file manifest will be required in which every file is associated with both the assay identifier and the sample identifier of the sample used to perform the assay and produce the file. The file manifest will provide the link between data file, sample, and assay. In addition, sample metadata, batch-level metadata, and quality control results will be required in the sample manifest. The two manifests together will provide all necessary metadata for FAIR representation of the data at time of data transfer and will allow for improved automation through data processing pipelines.

Planning phase effort can be minimized by creating a library of metadata templates for the most common assay methods from which the data planning team can select. The library object required for each assay can be linked to the assay repository such that when an assay is identified from the repository, the required metadata template is provided. The templates include the required and expected metadata definitions and format as well as instructions for the data-generating laboratory to transfer the data. If the required assay is not yet registered to the assay repository, it can be defined and registered as part of the planning phase, and the appropriate metadata template (existing or new) associated with it. The metadata library can be the basis for a series of tools to support the data planning and receiving teams. For any given instantiation of a library object for a specific dataset, a tool can be created to automate the data conformance to the standard, call the appropriate parsing or pipeline routines and set access controls.

7. CONCLUSION

7.1 FAIRification Effort

There is a business culture aspect to collecting data and metadata: in a broad curation project such as described here, many people across the organization provide input (documents, manifests, data locations, standards, etc.) to the process. Directive, advocacy and expectation from the top of the organization is highly effective for embedding a data culture throughout the organization. In the efforts described here, collaboration between the curation team and data owners is imperative to the success and efficiency of the process. When all parties agree on the value of the effort, processes can be developed across lines to meet the needs of the individual teams while also supporting FAIRification efforts across the organization. Engagement with the standards owners is also critical to success. Biomarker methods evolve very quickly, and new methods are regularly introduced. Standards for these methods and the metadata associated with them often lag behind the incorporation of the methods by clinical teams, resulting in a need to develop internal standards.

The majority of the work described in this paper addresses the findability and reusability of existing data collected over many years. The process is time-consuming, labor intensive, and aspects of metadata are often lost when not captured at the time the data are generated. This paper has described a semi-automated process to bring existing clinical trial biomarker data into FAIR recommendations. This process is particularly labor intensive due to the nature of how the existing data were collected in the absence of expectation for deep metadata. Several factors impact the effort required to curate a study, including the degree of change between the original version of standards and the version of standards to be applied, the number of technologies used to generate the biomarkers, and the total number of data sets to be curated. The curation team has provided a yearly average of approximately 10 Full Time Equivalent effort over 3 years thus far to this ongoing effort, resulting in over 700,000 samples mapped to more than 90,000 clinical trial participants and 4 million data files. Simultaneously, the team has curated 400 unique assays and assigned the resulting unique assay identifiers to more than 2 million files. While the output of the team grew faster than the size of the team (Figure 2) by improving processes and quality control review time with semi-automation, the effort reported here has curated approximately 25% of the total existing biomarker file repository. It is difficult to substantially improve the efficiency of this process when curating existing data collected under non-standard conditions, emphasizing the need for FAIR planning. The goal of the future planning process is to greatly improve this efficiency of creating FAIR data by collecting the metadata proactively and in a standard form.

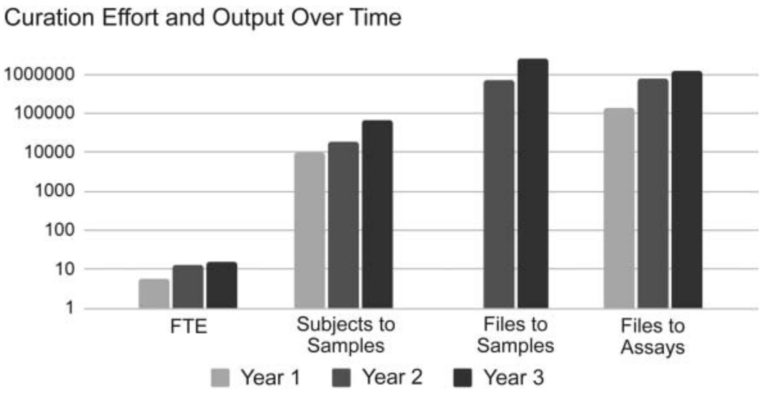

Figure 2. Curation effort and output. The total number for each year is provided for the following effort and output characteristics: Full Time Equivalents (FTE)—defined as one person for 40 hours per week for 50 weeks; Subjects to Samples—the number of unique subjects that have been mapped to at least one curated sample; File to Samples—the number of unique files that have been mapped to the sample(s) for which data were reported in the file; and Files to Assays—the number of unique files that have been mapped to the curated assay(s) that produced the reported data.

In view of the labor-intensive nature of the data FAIRification process described in this paper, it is reasonable to question if there are cost-saving alternatives to manual data curation and if there exists a fully automated data processing and curation solution, possibly leveraging the power of AI. Such fully automated solution has been considered and tried using one of the commercial data unification platforms that offered to leverage machine learning and other advanced algorithms to curate data at scale. Ultimately, only the semi-automated approach described in this paper achieved data FAIRification and integration at scale. Lack of consistency in data structure, content and format, data diversity (for each subset of data type, vendor, assay, etc.) and emergence of new types of data with time are among the contributing factors to the challenge of fully automating clinical and biomarker data curation. In our experience, there is no good alternative to the semi-automated process described above for the FAIRification of legacy data. As for the future studies, the manual component of the curation can be best minimized by addressing the inefficiencies in the business processes, rather than seeking advanced technical solutions. As described above in the Planning for FAIR Data section, planning and formalizing consistent data formats in advance of receiving data will significantly reduce costly manual curation efforts and speed up automation using scripted approaches developed in this work.

One aspect of FAIR, interoperability, has not been fully implemented in the current approach. In the efforts described here, harmonization of metadata has been achieved. However, the FAIR data principles advocate that the harmonized controlled terms be mapped to Universal Resource Identifiers (URI) where the identifier is stable and can be mapped to various synonyms from source ontologies and terminologies. Practically, the URI is often a complex construct and is not amenable to human interpretation of the data so that the URIs need to be converted back to one of the terms that the user community can interpret. This concept is important to provide different views of the data to different consumers, where terms preferred by one community may differ from another. The use and value of URIs has been well covered [12, 23, 24, 25] and the implementation of a terminology or ontology system to serve and translate URIs is out of scope of this current work.

7.2 FAIR Data Value

FAIRification of data greatly improves the ability to reuse data and by extension share it for further analysis either internally or externally. The processes described here bring metadata and data together in a way that only a single access point is needed to obtain data for analysis projects, whereas previously the information was scattered and disconnected. While this greatly improves overall accessibility to the data, individual access controls and sharing policies must be developed and maintained in concordance with local laws and study documents. Minimally, most clinical data sets will require access controls to conform to privacy and data sharing requirements associated with patient informed consent documents or other clinical study materials. Broader access can be appropriate with certain data anonymization practices, typically performed after the study is closed, cleaned and curated [26, 27, 28].

There are numerous options for sharing the data with the community. Choice of implementation should be based on user requirements. Some user communities will want to access processed data programmatically while others will require that data be presented as completed analyses. Importantly, the FAIRification process should prioritize community needs and several sharing environments can be implemented from the same data repository.

From a business perspective, the outcome of this effort has been broad and varied. The Subject-Sample mapping was the original process implemented and enabled the early integration of RNA-Seq data sets with the clinical attributes for several internal data marts. While allowing for appropriate access controls, these data marts facilitate the ability of data scientists to readily find the data and re-use it to address scientific questions that were previously considered too time-consuming or even unapproachable. Outcomes of these analyses include identification of potential new targets for cancer immunotherapy, and ability to quickly address product questions from government agencies. Additional access options, such as database interfaces and visualization tools, have been implemented where the biomarker scientist community has independently investigated these data sets for simple relationships such as the behavior of a biomarker in different patient populations, performance of different biomarker assays, or correlation between biomarkers and genetic markers. Previously, these activities would have required effort from the biostatisticians or data scientists to find, merge, clean and analyze the data, presenting a huge barrier to such exploration. These questions are now readily addressed, advancing research initiatives quickly and efficiently. In addition, more formal internal analytical “challenges” have leveraged these clean data sets to crowd-source analytical efforts to further scientific understanding, propose new treatment hypotheses, and drive new treatment development. The improved analytic capability achieved from the data marts is a direct result of implementing the FAIR principles: improving findability of data previously disconnected from the clinical record while applying robust access controls appropriate to the use and users of the data, which in turn support the re-use of the studies for more powerful analyses. Over time, the processes evolved and matured to include the additional assay metadata and relationships described above. The additional depth of metadata is now available through a series of tools that allow scientists to explore existing data to determine if data already exist to meet their needs, saving precious time and resources on the path to anticipating and achieving patients' needs.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the following individuals for their contributions to the overall strategy and vision of the project: Chris Morrow, Brian Sue, Steve Owens, Steven Day, and Wes Miyazaki on the development of an assay registration system referenced in this paper; Jerry Tang and Alina Bell for the development of a biomarker data repository referenced in this paper; Lala Motlhabi and Brandon Boylan for development of data processing and visualization processes that sit downstream of the work described in this paper; David Choy and Craig Cummings for helping define scientific use cases related to the work referenced in this paper; Jay Cowan for advice and guidance on data curation practices related to the work referenced in this paper; Kim Barnholt, Benjamin Szilagyi and Karen Petrie for helping to drive a Roche-wide initiative on FAIR data related to the work referenced in this paper; and Martin Romacker for the development of a terminology management system used by the team in the paper. The authors also wish to acknowledge Aruna Myneni for managing the project; Julie Bryant for providing inspiration and management; Amber Schedlbauer, Andrew Hill, Lada Sycheva, Liang-Tien Hsieh, Lucia Lam, Mangalalaxmy Subramanian, Marjolein Sponk, Susan Maskery, Timothy Warrington, Wei Deng, Whitney Dolan and Ying Wen for their data curation efforts; and Brandy Dykhuizen for editing the manuscript.

AUTHOR CONTRIBUTIONS

M. Whitley (maryann.whitley@ranchobiosciences.com) is the team lead for this project. She provided overall technical leadership, co-designed the conceptual data model, performed curation work and contributed to writing and editing the manuscript. M. McCreary (mccreary.mark@gene.com) defined the scope of the problem, co-designed the overall strategy, planned the overall detailed work with other project leads and contributed to writing and editing the manuscript as a senior author. A. Arefolov (alex.arefolov@ranchobiosciences.com) designed scripts and curation workflow for linking metadata for samples and data files, performed data curation, coordinated writing the manuscript and contributed to writing and editing the manuscript. H. Huang (huang.hongmei@gene.com) and C. Lu (lu.christina@gene.com) helped define the overall strategy and vision for the data management process, defined plans for informatics systems and provided project sponsorship. C. Farhy (chen.farhy@ranchobiosciences.com) performed data curation, designed many of the automated QC steps and contributed to writing and editing the manuscript. J. Wong (wong.jeeyeon@gene.com) and D. Das (das.diya@gene.com) contributed to building the technical process for data planning and acquisition and to defining requirements for curation workflows. Y. Budovskaya (yelena.budovskaya@ranchobiosciences.com) designed scripts and curation workflow for assigning assay metadata to data files and recoding assay information and performed data curation. K. Kanigel (kimberly.winner@ranchobiosciences.com), R. Ferguson (rebecca.ferguson@ranchobiosciences.com) and

O. Polesskaya (oksana.polesskaya@ranchobiosciences.com) designed scripts and curation workflow for assigning assay metadata to data files and recoding assay information, performed data curation and contributed to writing the manuscript. T. Staton (staton.tracy@gene.com) defined key scientific use cases and general problem statements in the data process and contributed requirements for the data management process. S. Brown (shoshana.brown@ranchobiosciences.com) and L. Adam (laura.adam@ranchobiosciences.com) performed data curation and contributed to writing the manuscript. X. Zeng (xiangpei.zeng@ranchobiosciences.com), R. Tajhya (rajeev.tajhya@ranchobiosciences.com) and C. Chen (cchen@championsoncology.com) designed scripts and curation workflow and performed data curation. All authors reviewed the manuscript.

REFERENCES

[1] Reinsel, D., Ganz, J., Rydning, J.: A digitization of the world: From edge to core. An IDC White Paper (2018). Available at: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. Accessed 29 June 2021

[2] Raghupathi, W., Raghupathi, V.: Big data analytics in heathcare: Promise and potential. Health Information Science and Systems 2(3), Article number 3 (2014)

[3] InsideBIGDATA Guide to Healthcare & Life Sciences (2016). Available at: https://insidebigdata.com/2016/09/27/insidebigdata-guide-to-healthcare-life-sciences/. Accessed 29 June 2021

[4] Wilkinson, M., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018 (2016)

[5] FAIR Principles. Available at: https://www.go-fair.org/fair-principles/. Accessed 29 June 2021

[6] G7 Expert Group on Open Science. Executive Summary (2017). Available at: http://www.g8.utoronto.ca/science/2017-annex4-open-science.html. Accessed 29 June 2021

[7] NIH Data Commons Pilot Phase Consortium (2018). Available at: https://commonfund.nih.gov/commons/awardees. Accessed 29 June 2021

[8] Turning fair into reality: Final report and action plan from the European Commission Expert Group on Fair Data (2018). Available at: https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf. Accessed 29 June 2021

[9] Staines, R.: Pfizer follows Novartis and GlaxoSmithKline by appointing new Chief Digital Officer (2018). Available at: https://www.healthcare.digital/single-post/2018/10/10/Pfizer-follows-Novartis-and-GlaxoSmithKline-by-appointing-new-Chief-Digital-Officer. Accessed 29 June 2021

[10] Digital innovation strategy for Roche. Available at: https://www.boardofinnovation.com/client_cases/digital-innovation-strategy-for-roche/. Accessed 29 June 2021

[11] Chan, H.C.S., et al.: Advancing drug discovery via artificial intelligence. Trends in Pharmacological Sciences 40(8), 592–604 (2019)

[12] Wise, J., et al.: Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discovery Today 24(4), 933–938 (2019)

[13] Vadas, A., Bilodeau, T.J.: The evolution of biomarker use in clinical trials for cancer treatments. L.E.K. Special Report (2019). Available at: https://www.lek.com/insights/sr/evolution-biomarker-use-clinical-trials-cancer-treatments. Accessed 29 June 2021

[14] Carini C., Fidock M., D., van Gool A., J. (eds): Handbook of biomarkers and precision medicine. 1st edition. Chapman and Hall/CRC, Boca Raton (2019)

[15] Dakappagari, N., et al.: Application of biomarkers in oncology clinical trials. Clinical Investigation 5(1), 61–74 (2015)

[16] Thomas, D.W., et al.: Clinical development success rates 2006–2015. A BIO Industry Analysis White Paper (2016)

[17] U.S. Food and Drug Administration: Study data technical conformance guide. A Technical Specifications Document (2018). Available at: https://www.fda.gov/media/88173/download. Accessed 29 June 2021

[18] Izard, D.: Preparing legacy format data for submission to the FDA: When & why must I do it, what format should I follow? PharmaSug paper (2016). Available at: https://www.pharmasug.org/proceedings/2016/SS/PharmaSUG-2016-SS02.pdf. Accessed 29 June 2021

[19] Mohanty, S., et al: The development and deployment of Common Data Elements for tissue banks for translational research in cancer – An emerging standard based approach for the Mesothelioma Virtual Tissue Bank. BMC Cancer 8(91), Article number 91 (2008)

[20] Mirbagheri, E., Ahmadi, M., Salmanian, S.: Common data elements of breast cancer for research databases: A systematic review. Family Medicine and Primary Care 9(3), 1296–1301 (2020)

[21] Badawy, R., et al.: Metadata concepts for advancing the use of digital technologies in clinical research. Digital Biomarkers 3, 116–132 (2019)

[22] Wickam, H.: Tidy data. Journal of Statistical Software 59(10), 1–23 (2014)

[23] Wise, J., et al.: Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discovery Today 24(4), 933–938 (2019)

[24] Djokic-Petrovic, et al: PIBAS FedSPARQL: A web-based platform for integration and exploration of bioinformatics data sets. Journal of Biomedical Semantics 8, 42 (2017)

[25] Smith, J.R., et al.: The clinical measurement, measurement method and experimental condition ontologies: Expansion, improvements and new applications. Journal of Biomedical Semantics 4(1), 26 (2013)

[26] Chevrier, R., et al.: Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review. Journal of Medical Internet Research 21(5), e13484 (2019)

[27] Kayaalp, M.: Patient privacy in the era of big data. Balkan Medical Journal 35(1), 8–17 (2018)

[28] Kayaalp, M., et al.: Challenges and insights in using HIPAA Privacy Rule for clinical text annotation. In: AMIA Annual Symposium proceedings, pp. 707–716 (2015)

AUTHOR BIOGRAPHY

Alexander Arefolov is a Senior Data Scientist at Rancho Biosciences, currently leading curated data integration and loading effort within cross-functional Rancho-Genentech data management team. Since joining Rancho, he has worked in support of automating data management, data curation and data FAIRification projects. Before that, Alexander worked for over 15 years in pharmaceutical industry and academia as an organic and medicinal chemist, drug discovery scientist and consultant. He has a strong interest in applications of AI and machine learning in healthcare. Alexander earned MS in chemistry from Moscow State University, Moscow, Russia. He completed his PhD in chemistry at Boston University in the area of natural product synthesis. Postdoctoral training was completed at Harvard University in the area of total synthesis and drug design.

ORCID: 0000-0002-3447-3223

Laura Adam is a Data Scientist at Rancho Biosciences. Prior to working on clinical data FAIRification, she had been developing software tools and knowledge graphs for synthetic biologists for design, construction, and analysis of genetic constructs. She consulted US governmental agencies on bioinformatics solutions to mitigate biosecurity risks, as well as participated in the development of the synthetic biology open language (SBOL).

ORCID: 0000-0002-4822-2695

Shoshana Brown is a Senior Scientist at Rancho Biosciences. She holds a bachelor's degree in Biochemistry and Molecular Biology from the University of California, Santa Cruz. Her PhD work and subsequent research at the University of California, San Francisco focused on enzyme sequence-structure-function relationships in large and diverse enzyme superfamilies using similarity networks. She is currently a team leader at Rancho Biosciences, where she is involved in the curation of clinical and exploratory biomarker data to SDTM and other formats, with a focus on automating curation workflows whenever possible.

ORCID: 0000-0002-0875-4902

Yelena Budovskaya (PhD, Ohio State University) is a Biomedical Data Curation specialist with more than 20 years of R&D expertise in the biology of aging, cancer genetics, genomics, and microbiology. Currently she creates and executes strategy around all areas of the data lifecycle management, data standards, and ongoing data operations for Genentech, A Member of the Roche Group, ensuring that all clinical data follow the FAIR data principles.
ORCID: 0000-0002-1047-693X

Cong Chen is a Computational Biologist, who is currently working for Champions Oncology Inc. Since receiving his PhD degree in Genomics in 2005 from Beijing Genome Institute (CAS), Cong Chen has contributed to drug discovery field through his roles in both academia and industry. He finished his Post-doctoral training at Northwestern University, where he published findings in *Nature Immunology* and other high-profile journals. He is interested is in big data, deep learning and using those techniques to expedite the drug discovery process.
ORCID: 0000-0003-0907-0374

Diya Das leads data planning and data sharing efforts for the Development Sciences Informatics Data Management Team at Genentech, where she is a Senior Informatics Analyst. She received her PhD in Molecular & Cell Biology from the University of California, Berkeley, where she was a Moore-Sloan Fellow at the Berkeley Institute for Data Science. She received an AB in Molecular Biology and Certificate in Neuroscience from Princeton University.
ORCID: 0000-0001-9646-8983

Chen Farhy is a Senior Scientist with Rancho Biosciences with a strong background in computer sciences and automated image analysis. He received his PhD degree from the University of Tel-Aviv for his work analyzing the genetic networks that regulate retinal progenitor cell proliferation and differentiation during mammalian eye development. For his postdoctoral training at Sanford Burnham Prebys Medical Discovery Institute he developed novel image based profiling assays for high content screening with a focus on glioblastoma stem cells differentiation and epigenetic modulation.
ORCID: 0000-0001-6160-3479

Rebecca Ferguson began her scientific journey as an undergraduate double major, earning Bachelor degrees in both Biology and Zoology at Colorado State University. She continued her training at the University of Colorado Anschutz Medical Campus, earning a Molecular Biology PhD in 2009 under the guidance of Dr. James Maller, investigating regulatory mechanisms linking centrosome duplication and DNA replication. Her postdoctoral work with Dr. Robert Sclafani expanded on these investigations to define the functional helicase required for DNA origin initiation and mechanisms for DNA damage bypass during active replication. After nearly a decade at the bench as a laboratory and clinical trial site manager, Dr. Ferguson decided to step out of the laboratory and now works as a scientist and data curator. By reorganizing and structuring clinical data with the goal of ensuring maximum use and discovery, Dr. Ferguson strives to identify new options and treatments for better patient care and survival.
ORCID: 0000-0001-9528-2466

As the Vice President of Development Sciences Informatics at Roche Genentech, **Hongmei Huang** is responsible for the strategic leadership around the data ecosystem and AI/Analytics platforms for translational sciences. She is among the key leaders driving the Roche wide effort to make our data FAIR. By connecting science and technology, Hongmei leads organizational drives to transform the data and digital landscape for the advancement of medicines and healthcare. She is an accomplished scientific and informatics leader with over 25 years of experience in the Pharmaceutical Industry. She started her career as a Research Investigator at Bristol-Myers Squibb and transitioned into Informatics over the course of her career, with leadership roles in various companies including Novartis and Johnson & Johnson. Hongmei received her B.S. from Beijing University, M.S. from University of Michigan, and PhD in BioOrganic Chemistry from The Scripps Research Institute.
ORCID: 0000-0002-6828-7365

Kimberly Kanigel Winner is a Scientist at Rancho Biosciences. She completed her postdoctoral training under an NIH training grant in the Computational Biosciences Program at University of Colorado School of Medicine. Her PhD in Computational Biology was completed at The University of New Mexico and the New Mexico Center for the Spatiotemporal Modeling of Cell Signaling. Her Bachelor's degree in Biology/Chemistry minor was earned at Fort Lewis College. Her primary background is in two- and three-dimensional modeling of drug delivery in tumors, as well as RNAseq analysis and other bioinformatics, and high performance computing support. She is currently a project technical leader for biological assay metadata mapping at Rancho Biosciences.
ORCID: 0000-0002-6044-838X

Christina Lu is the Sr. Director of the Data Management and Engineering team in DevSci Informatics. She is an accomplished informatics leader with more than 20 years of experience in the Biotech and Pharmaceutical Industries. She has experience working from Research to Development environments, leading the strategy and implementation of informatics solutions and data platforms. At Genentech, Christina and her team developed a comprehensive strategy for managing data from data ingestion to analysis, and building data management processes and data ecosystems that are flexible and scalable to handle new data types. Christina is also a core team member of the EDIS program, a cross-Roche initiative focusing on making data FAIR prospectively and retrospectively to maximize the value of data. Prior to joining Genentech, she was at Roche Palo Alto, Exelixis and more recently Novartis, where she was the Director and Site Head leading the informatics and engineering teams based in Emeryville. Christina received her M.S. in Biomedical Informatics from Stanford University. She also has a M.S. in Immunology and B.S. in Biochemistry/Cell Biology from UC San Diego.
ORCID: 0000-0001-9911-3141

Oksana Polesskaya has over 20 years of experience in basic and translational science. Oksana has expertise in neurobiology, immunology and genetics and has deep understanding of methodology in broad range of assays, from histology to immunoassays to a variety of NGS methods. She co-authored 27 research papers in peer-reviewed journals. Oksana Polesskaya received MS from Moscow State University, and PhD degree from the Institute of Medical Genetics in Russia. Oksana currently works as a Research Scientist at University of California San Diego.
ORCID: 0000-0003-3024-114X

Tracy Staton completed her undergraduate training at UC Berkeley and her graduate studies in the Immunology Program at Stanford University. At Stanford, she worked in the laboratory of Dr. Eugene Butcher to characterize molecular mechanisms of lymphocyte homing. As a postdoc in the laboratory of Dr. Laurie Glimcher at Harvard, she studied immune cell development and the importance of effector lymphocyte populations in preclinical models of disease. As part of the OMNI-Biomarker Development department at Genentech, Tracy developed biomarker strategies for respiratory indications (including asthma and COPD).
ORCID: 0000-0002-2718-9638

Rajeev Tajhya is a Scientist and Technical Lead at Rancho Biosciences. He received his PhD with a focus on electrophysiology of ion channels in diseases from Baylor College of Medicine, Houston, TX. He worked on ion channels in embryo development during his postdoctoral training at University of California SF. Rajeev is drawn to big data in human physiology and fascinated by the advancements in data science. He is interested in application of data science tools to finding and automating solutions with data.
ORCID: 0000-0002-5363-2293

Maryann Whitley earned a B.S. in clinical laboratory science from University of Florida. She completed her PhD in Molecular Immunology at Harvard University School of Public Health. Postdoctoral training was completed at Brigham and Women's Hospital in the area of transcriptional regulation of endothelial cell activation. She has over 20 years working in the biotech/pharma industry. While contributing to the validation of lab methods for transcriptional profiling with Affymetrix GeneChips, she quickly realized her strength in data manipulation and analysis including the development of early data normalization concepts. She moved from the lab to computational biology role with primary responsibility for analysis of transcriptional profiling data, eventually leading a staff of 3-5 computational biologists analyzing genomic scale data for hundreds of researchers. Since joining Rancho Biosciences, she has led a multi-disciplinary team in a large-scale effort to FAIRify clinical and biomarker data. Her areas of strength include immunology, oncology, transcriptional regulation, and many aspects of clinical research data. Her technical skills include SQL, R, and many commercially available software tools.
ORCID: 0000-0001-5228-1113

Jee-Yeon Wong is a Senior Informatics Analyst in Development Sciences Informatics at Genentech. She is responsible for leading the Data Acquisition team within the Data Management team. Under her leadership, her team is responsible for the acquisition of data from external parties to support biomarker clinical studies. Prior to her current role, she spent many years as a principal informatics analyst in research informatics. Some of her highlighted achievements were developing a small molecule sample inventory management and request system, research assay registration systems, and managing the research biological databases. She holds a Bachelor's in Zoology and Economics from UC Davis.
ORCID: 0000-0002-8201-1904

Xiangpei Zeng is a Senior Scientist at Rancho Biosciences. He completed his postdoctoral training and received his PhD in University of North Texas Health Science Center. His Master and Bachelor degrees in forensic genetics were earned at Sun Yat-sen University, China. His primary background is selecting highly informative STR and SNPs for human ancestry estimation and human identification. He is currently leading a data curation team at Rancho Biosciences.
ORCID: 0000-0003-1861-5903

Mark McCreary is the Director of DevSci Data Management and Curation in Development Sciences Informatics at Genentech. His team focuses on managing exploratory and high-dimensional data on behalf of clinical development and translational scientific teams. In his role, he has led efforts to curate data assets for high-priority therapeutic areas; standardized sample and assay metadata; and developed streamlined end-to-end data workflows from planning through processing that yield FAIR data and enable exciting analytics. Mark has a strong interest in the intersection between biology, medicine, and computational biology, and has been exposed to multiple research areas. He joined Genentech in 2015, focusing on Bioinformatics efforts related to Immunology and Infectious Disease Research efforts. He moved from an analytics role to data management as he grew interested in streamlining data integration for analytical use, ultimately driven by inefficiencies he saw that are often faced by computational scientists when preparing data for analytical activities. Prior to his engagement with Genentech, Mark worked as a Bioinformatics Analyst in Stan Cohen's laboratory at Stanford University. His informatics expertise includes data lifecycle management; gene expression and pathway analysis; and application development in R, Perl, and Python. He holds a Masters in Bioinformatics from RIT.
ORCID: 0000-0002-5590-7748